The 2015 Language Recognition

i-Vector Machine Learning Challenge

1 INTRODUCTION

The National Institute of Standards and Technology (NIST) will coordinate a special i-vector challenge in 2015 based on data used in previous NIST Language Recognition Evaluations (LREs)[1] and certain other sources. This challenge is intended to foster interest in this field from the broader machine learning community. It will be based on the *i-vector* paradigm widely used by state-of-the-art speaker and language recognition systems, and will largely follow the approach taken in the recent NIST-coordinated speaker recognition i-Vector Challenge.[2] By providing i-vectors directly, and not utilizing audio data, the evaluation is intended to be readily accessible to participants from outside the audio processing field.

The i-vectors supplied will be based on a system developed by the Johns Hopkins University Human Language Technology Center of Excellence in conjunction with MIT Lincoln Laboratory.[3]

Registered participants may offer multiple challenge submissions. (A limit will be established of 100 per day per registered participant). A leaderboard will be maintained by NIST indicating the best submission performance results thus far received and processed.

2 TECHNICAL OBJECTIVE

This challenge focuses on the development of new methods for using i-vectors for language identification in the context of conversational telephone or narrowband broadcast speech. It is designed to foster research progress, including goals of:

- Exploring new ideas in machine learning for use in language recognition
- Making the language recognition field accessible to more participants from the machine learning community
- Improving the performance of language recognition technology.

3 TASK

The challenge focuses on the task of **language identification.** This task is to determine which, if any, language from a set of possible *target languages* is being spoken in a given *test segment* of conversational speech. This challenge will be offered in an *open set* context, meaning that the actual language of the test segment may not be any of the specified target languages.

The challenge will consist of a defined set of 50 target languages and of a sequence of 6500 test segments, each specified by a single i-vector. The target languages will be ones used previously in NIST LRE's, and for each a set of approximately 300 training segment i-vectors will be provided.

For each test segment i-vector, the system will be required to specify the name of the language class to which it is believed to correspond, or "out_of_set" if the segment is believed to correspond to none of the target languages.

4 PERFORMANCE MEASURE

The primary scoring metric, to be used in maintaining the challenge leaderboard, is defined as follows:

$$Cost = \frac{(1 - P_{OOS})}{n} * \sum_{k}^{n} P_{error}(k) + P_{OOS} * P_{error}(oos)$$

$$\text{where } P_{error} = \left(\frac{\#errors\_class\_k}{\#trials\_class\_k}\right), n = 50, \text{and } P_{OOS} = 0.23$$

Results will also be examined where test segments are limited to those where "out_of_set" is not the correct response (the *closed set* case), and for other subsets representing conditions of particular interest.

Thus for each challenge participant, the primary performance score returned for a system submission during the challenge will be the computed cost function over all test segments used for progress scoring (see the test segment subset division discussion in section 5.5). The leaderboard will be ordered by these best identification progress scores. Other scores of interest may be displayed as well. At the conclusion of the challenge, the score for a site's final submission will be determined based on the evaluation set of test segments.

5 DATA

The input data provided will consist of development data to be used for system creation, language training data for each target language, and test segment data. It may be noted that every attempt was made to use any given speaker only once in all of the data.

The i-vectors provided will be derived from conversational telephone and narrowband broadcast speech data utilized for the NIST Language Recognitions Evaluations (LRE's) from 1996 to 2011 and from certain other sources, including the IARPA BABEL Program[4]. Each i-vector will be a vector of 400 components. Along with each i-vector, a single item of metadata will be supplied, namely the estimated duration of speech (in seconds) used to compute the i-vector. Segment durations are being chosen to have a log-normal distribution with a mean of approximately 35s.

5.1 Development Data

A set of about 6500, unlabeled i-vectors will be provided for general system development purposes. These will be from

[1] See http://nist.gov/itl/iad/mig/lre.cfm

[2] See https://ivectorchallenge.nist.gov/

[3] See N Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language Recognition via Ivectors and Dimensionality Reduction", *Proc. Interspeech 2011*, Florence, Italy, Aug. 2011

[4] Data used includes BABEL data from babel103b-v0.4b, babel101b-v0.4c, babel201b-v0.2b, IARPA-babel203b-v3.1a, babel104b-v0.4bY, babel106b-v0.2g, babel105b-v0.4, babel107b-v0.7, babel206b-v0.1d

segments in unspecified languages and may be used, for example, for unsupervised clustering in order to learn about wanted and unwanted variability in the i-vector space. These unspecified languages will include all of the target languages and multiple additional "out_of_set" languages.

5.2 Language Training Data

This will consist of a set of 300 i-vectors for each of the 50 specified target languages.

5.3 Test Segment Data

This will consist of 6500 single i-vectors representing test segments. These will include segments corresponding to all of the target languages and (an unspecified number of) segments involving various out_of_set languages.

5.4 System Output

The required outputs for each challenge submission will consist of the name of the target language (or "out_of_set") for each test segment i-vector. Thus the total number of such outputs will be 6500.

5.5 Test Segment Subset Division

The test segment i-vectors will be divided into two subsets: a *progress subset*, and an *evaluation subset*. The progress subset will comprise a randomly selected 30% of these, and will be used to monitor progress on the leaderboard. The remaining 70% of these will form the evaluation subset, and will be used to generate the official final scores determined at the end of the challenge. These subsets will not be made known to systems.

6 BASELINE SYSTEM

A baseline system will be included in the download package. This system will serve as an example of how to achieve a successful submission. Also, the performance of the system will be the baseline for the scoreboard. The algorithm used in the baseline is a variant of cosine scoring with the following recipe:

1. Use the unlabeled development data to estimate a global mean and covariance.
2. Center and whiten the evaluation i-vectors based on the computed mean and variance.
3. Project all the i-vectors onto the unit sphere.
4. For each language, average its training i-vectors and then project the resulting average-language i-vector onto the unit sphere.
5. Compute the inner product between all the average-language i-vectors and test i-vectors.

7 RULES

Each participant must complete the online registration process and download the development data i-vectors, and the target language and test segment i-vectors.

Each uploaded system submission must contain an output (target language designation) for all test segment i-vectors in order to be scored.

The output produced for each test segment i-vector may not use in any way the i-vectors for other test segments. Normalization over multiple test segments is not allowed.

Participants may report on their own performance in the challenge, but may not make advertising claims about winning the challenge or claim NIST endorsement of their system(s). The following language in the U.S. Code of Federal Regulations (15 C.F.R. § 200.113) shall be respected[5]:

> *NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST, or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.*

8 SYSTEM DESCRIPTION

Each participant is asked to provide an overall description of the algorithm and procedures used to create the submitted systems, which may be shared with the community. Please send system descriptions to lang_ivector_poc@nist.gov.

9 SCHEDULE

April 20:	Registration opens on website
May 15:	Challenge data available on website
September 1:	Last day to submit output for official scoring
September 2:	Official scores (on evaluation subset) posted

[5] See http://www.ecfr.gov/cgi-bin/text-idx?c=ecfr&rgn=div5&view=text&n16-Apr-15ode=15:1.2.2.1.1&idno=15#15:1.2.2.1.1.0.21.14